PINYIN CONVERSION SPECIFICATIONS

CONVERSION SEQUENCE JK: JAPANESE/KOREAN DISCRIMINATOR PATTERNS

FINAL VERSION, OCTOBER 12, 2001

AY == ayn or apostrophe or alif

1.1 The purpose of this conversion sequence is to identify subfields which could not be Wade Giles, despite any dictionary matches, and to lessen the risk of converting Korean or Japanese subfields whose syllables all match Wade Giles syllables.

1.2 JAPANESE -- PATTERNS THAT WILL NEVER OCCUR IN WADE-GILES SYLLABLES

1.2.1 Initial b, d, g, r, z (i.e. ba, da, ga, ra, za) [valid, but generates too much noise]; second letter y (i.e. kyo, ryu, pyan)

1.2.2 Restrict to the following larger matches: kyo, ryu, pyan;

1.2.3 Final aa, ae, au, ea, ee, eo, eu, ie, ii, io, oa, oe, oi, oo, on, ue, uu;

1.2.4 Syllables with macrons over the letters o or u

1.3 KOREAN -- PATTERNS THAT WILL NEVER OCCUR IN WADE-GILES SYLLABLES

1.3.1 Initial letters

1.3.1.1 r [valid, but generates too much noise]

1.3.1.2 Initial patterns:

Four character patterns: ch<AY>w Ch<AY>w, ch<AY>y, Ch<AY>y;

Three character patterns: chw, Chy, chy, Chy, k<AY>w, K<AY>w,k<AY>y, K<AY>y, p<AY>w, P<AY>w, p<AY>y, P<AY>y, T<AY>w, T<AY>w, t<AY>y, T<AY>y;

Two letter patterns: hw,Hw,hy,Hy, kw,Kw,ky,Ky, mw, Mw, my, My, nw, Nw, ny,Ny, pw,Pw, py,Py,ry,Ry,sw,Sw, tw,Tw

1.3.1.3 Double initial consonants: pp, tt, kk, tch [use if needed?]

1.3.2 Medial letter w or y, as limited below: Restrict "medial w" pattern matches to the following strings of initial patterns: kw ,pw ,tw ,chw ,nw ,mw ,sw , shw, hw, k'w ,p'w ,t'w ,ch'w

Restrict "medial y" pattern matches to the following strings: ky, py, chy, ny ,ry ,my , shy , hy , k'y, p'y, t'y, ch'y

1.3.3 Final p, t, k, m, l [too broad]

1.3.4 Syllables with breves over the letters o or u

1.4 Application of the Patterns and Use of Dictionaries

1.4.1 Flagging for Japanese or Korean: check the language code for Japanese or Korean, check the 041, and check the 546 for the text "Japanese" or "Korean" in determining whether to set the flag for a given record. In cases where the flag is set, do searches of the dictionaries and multicharacter patterns noted below to guard against converting Japanese.

1.4.2 Diacritics checking:

1.4.2.1 Screen all subfields, regardless of whether the record is coded Japanese or Korean, for the breve-o, breve-u, macron-o and macron-u.

1.4.2.2 If found skip the subfield. In Mixed Text subfields, skip the subpart of the subfield containing the match, and evaluate the rest of the subfield as usual under Mixed Text processing.

1.4.3 Dictionaries:

1.4.3.1 Use the Japanese WG or Korean WG dictionary only if the record is coded Japanese or Korean or has that language code, but not if a place names table match has been found [patterns might overlap].

1.4.3.2 If a subfield has WadeGiles, Same or Common syllables search the appropriate dictionary for common matches in the Japanese or Korean dictionary. Count each syllable that is also matched in the Japanese or Korean dictionary.

1.4.3.3 Tally the total of UWG/Same/Common syllables that are also found in the dictionary. If they are equal in number, and there is "Other" text of unknown type, skip the subfield. If the tallies are equal in number, and there is no discriminating text , flag the subfield as Review Required.

1.4.4 Multicharacter Pattern checking: look for patterns of characters found in Japanese or Korean that would never occur in Wade Giles, and so might help to exclude the subfield.

1.4.4.1 Use these patterns where Japanese or Korean is indicated by the 041 or by the language code, or the 546.

1.4.4.2 Screen for Japanese final two letter combinations and use this check to resolve unknown text.

1.4.4.3 Screen for initial 4-letter, 3-letter, 2-letter patterns, longest first, to resolve unknown text.

1.4.4.4 Other possible Japanese/Korean “never in Wade Giles” patterns as found useful in the course of testing.